Streamlining Data Exploration and Visualization in Climate Science

E. Santos[1], H. Vo[1], T. Fogal[1], D. Williams[2], C. Doutriaux[2], and C. Silva[1]

[1]Scientific Computing and Imaging Institute, University of Utah; [2]Lawrence Livermore National Laboratory

Summary

The amount of data generated by global atmospheric and oceanic models is enormous. To help climate researchers make sense of all this data, we are working together with the *Earth System Grid Center for Enabling Technologies (ESG-CET)* to provide these researchers with access to data, information, models, analysis, visualization tools, and computational resources. Climate scientists will be able to efficiently perform comparative analysis and visualizations of their experiments, by leveraging a scientific workflow and provenance management system and climate and data analysis tools in a highly collaborative problem-solving environment.

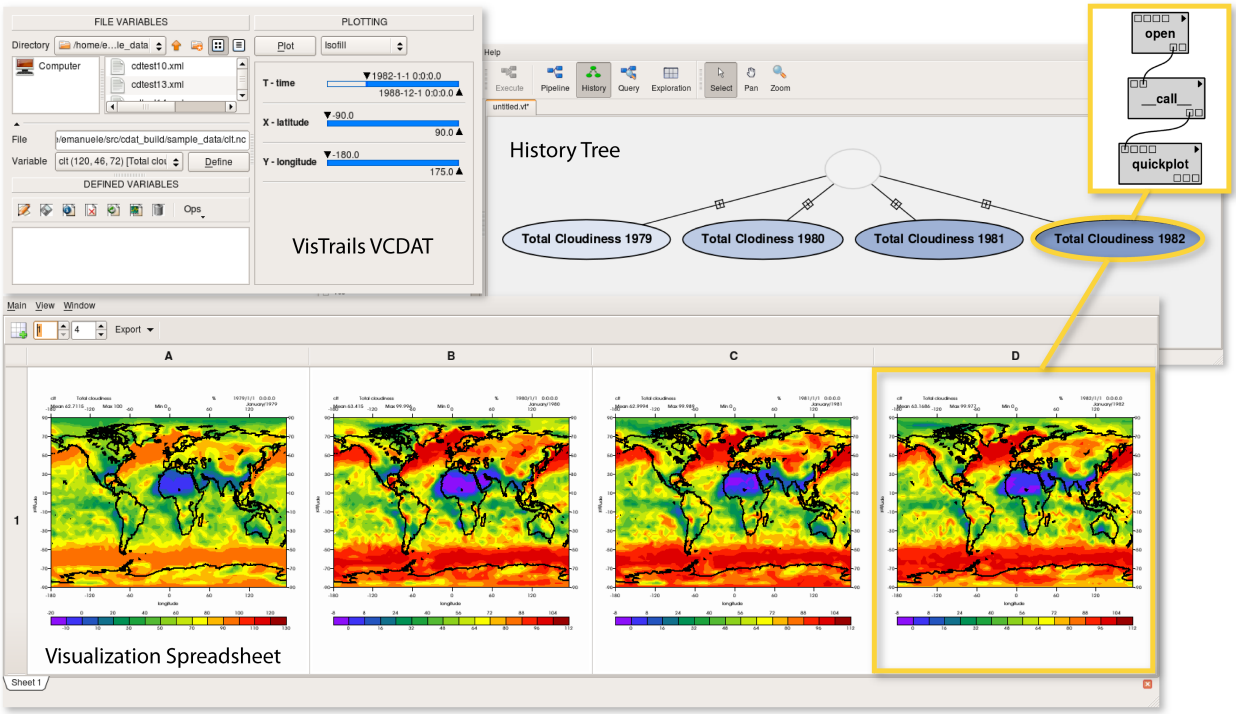

Figure 1. Example of comparative visualization using the CDAT VisTrails package. Using VisTrails parameter exploration mechanism and the visualization spreadsheet, a user quickly generates a side-by-side comparison of total cloudiness in the month of January from 1979 to 1982. As the user interacts with the VisTrails VCDAT window to select and manipulate variables in the climate datasets, VisTrails automatically generates and captures provenance information. This information reflects the "recipe" for creating a given visualization or series of visualizations. The "recipe" can be shared, reused, or even compared with other "recipes" to reveal differences and similarities in different types of data analysis.

One of the most difficult challenges that climate scientists face today is managing and understanding the massive amounts of global atmospheric and oceanic model data collected by observation and generated by simulations. In order to facilitate data processing, management, retrieval and analysis, the *Program for Climate Model Diagnosis and Intercomparison (PCMDI)* and other DOE sites started the *Earth System Grid (ESG)* project, which allows users anywhere to remotely access a distributed multi-petabyte archive and perform analysis of climate datasets, which are available on supercomputers and large-scale data and analysis servers. To meet its data

access and analysis needs, PCMDI also developed the *Climate Data Analysis Tools (CDAT)*, a software system that leverages the ESG infrastructure to support exploration and visualization of climate scientific datasets.

The climate community widely adopted CDAT for accessing and analyzing data in their scientific experiments, but as data size and complexity has grown, climate researchers are faced with new challenges. First, the amount of data resulting from simulations has grown too large to be easily moved across the network, making it more difficult to be analyzed and visualized. Second, scientists frequently need to compare results of simulations. For example, when they are trying to understand the effect of parameter changes on resulting simulation output. These comparisons might generate hundreds of different configurations of datasets, parameter values and visualization techniques. Keeping track of all this information using traditional (pen and paper) methods is time-consuming and error-prone.

Our team is working with the ESG-CET to address both of these challenges. The ultimate goal is to provide climate researchers worldwide with access to: data, information, models, analysis, visualization tools, and computational resources required to make sense of enormous climate simulation datasets. We started adding provenance support (see "Provenance and its Importance for Science" sidebar) and comparative visualization tools to CDAT, by integrating it to the VisTrails system. VisTrails is a scientific workflow and provenance management system developed at the University of Utah that provides support for data exploration and visualization.

The result is a CDAT VisTrails package that allows users to leverage functionality from CDAT and VisTrails, including the capability to perform comparative visualizations, while keeping the entire provenance generated during the data exploration and visualization (see Figure 1). Using a graphical interface similar to the one already available in CDAT, users can use familiar tools to build their workflows and have their entire provenance automatically captured.

Provenance and its Importance for Science

Provenance is defined by the Oxford English Dictionary as the source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners. Provenance is important for science because it helps to interpret and reproduce the results of an experiment; to understand the chain of reasoning used in the production of a result; to verify that the experiment was performed according to acceptable procedures, to track who performed the experiment and who is responsible for its results.

The goal is to speed up the scientific discovery process, allowing scientists to easily examine all the steps that led to a result, identify the experiment's inputs and outputs and consequently help reproduce the results.

VisTrails brings functionality that supports concurrent exploration of multiple visualizations with the use of a spreadsheet (see Figure 1), which together with the parameter exploration mechanism allows users to effectively compare visualizations produced by different workflows side by side.

By collaborating with the PCMDI and the ESG-CET, VisTrails will be part of an overall solution to facilitate data to a worldwide audience of climate data consumers. Together, this integrated enterprise system is designed to manage and analyze extremely large and diverse datasets.

For further information on this subject contact:
Name: Cláudio T. Silva.
Organization: Scientific Computing and Imaging Institute, University of Utah.
Email: csilva@sci.utah.edu
Phone: (801) 587-7588

Note: This work was funded in part by the SciDAC2 Visualization and Analytics Center for Enabling Technologies and ASCR's Visualization Base Program by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.